Discussion on article “Bayesian inference with misspecified models” by Stephen G. Walker

Journal of Statistical Planning and Inference ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Discussion on article "Bayesian inference with misspecified models" by Stephen G. Walker

Pierpaolo De Blasi [1]

University of Torino and Collegio Carlo Alberto, Italy

The paper by Stephen Walker offers an interesting view of the rationale of Bayesian inference with misspecified models. The author resorts to the representation theorem of de Finetti to justify a more flexible use of Bayes theorem, flexible in that it requires less assumptions on the data generating process. Predictive densities are seen as guesses obeying some form of symmetry when learning from past observations. They can be chosen to define an exchangeable law for the observables which does not need to conform to the way the data are actually generated. Through the representation theorem, it is possible to separate the statistical model (that is, the likelihood) from the prior and look at the former as a suitable approximation for the stochastic phenomena of interest. Posterior inference has then to be validated in terms of its asymptotic behavior with respect to the data generating process.

In this note we would like to address two related issues. In Section 1 some results on the predictive construction of parametric models are reviewed; they help to gain some insight on Walker's use of Bayes theorem in case of misspecification. In Section 2 a parametric family of densities is considered. Given that the interest is in finding through posterior inference the parameter value that minimizes the divergence with the true density, it is worth to consider estimation of the minimum divergence with Bayesian nonparametric methods.

1. Predictive model representation

In the predictive approach to Bayesian inference, the model is defined as a predictive probability specification of the observables. The representation theorem provides a basis for separating out two components: a statistical model and a prior distribution for the parameter of interest. Both are characterized by the convergence of predictive distributions depending on a predictive sufficient statistic, where convergence is defined with respect to the exchangeable law they induce. Below we formalize these ideas without going into measure-theoretic technicalities.

Let $(X_n)_{n\geq 1}$ be a sequence of exchangeable random variables and its law be denoted by P. Also, let $\mathcal{F}$ be the space of all distributions on the real line $\mathbb{R}$ and $\hat{F}_n(\cdot)=\sum_{i=1}^n \delta_{X_i}(\cdot)/n$ be the empirical distribution of $X_1,\ldots,X_n$. Occasionally, we will use $X_{1:n}$ as short hand notation for $X_1,\ldots,X_n$. According to de Finetti representation theorem, there exists a unique probability measure μ on $\mathcal{F}$ such that, for any $n\geq 1$,

$$P(X_1\leq x_1,\ldots,X_n\leq x_n)=\int_{\mathcal{F}}\prod_{i=1}^n F(x_i)\,\mathrm{d}\mu(F). \tag{1}$$

Moreover, P almost surely (a.s.), $\hat{F}_n$ converges weakly to a random $\tilde{F}$ with law μ. The exchangeable law P defines a sequence of predictive distributions $(P_n)_{n\geq 1}$, where P_n is the conditional distribution of X_{n+1} given $X_1,\ldots,X_n$. In the predictive approach, the aim is at constructing an exchangeable law P by starting with a sequence $(P_n)_{n\geq 1}$. These predictive distributions have to satisfy, P-a.s., the following two conditions:

E-mail address: pierpaolo.deblasi@unito.it

[1] This work was supported by the European Research Council (ERC) through StG "N-BNP" 306406.

0378-3758/$ - see front matter
http://dx.doi.org/10.1016/j.jspi.2013.05.015

Please cite this article as: De Blasi, P., Discussion on article "Bayesian inference with misspecified models" by Stephen G. Walker. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.05.015

(a) $P_n(A|x_{1:n}) = P_n(A|x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ for any permutation $\sigma(1), \ldots, \sigma(n)$ of $\{1, \ldots, n\}$ and for any $n \geq 2$;
(b) $\int_B P_{n+1}(A|x_{1:n+1})\, \mathrm{d}P_n(x_{n+1}|x_{1:n}) = \int_A P_{n+1}(B|x_{1:n+1})\, \mathrm{d}P_n(x_{n+1}|x_{1:n})$ for every $A, B \subset \mathbb{R}$ and any $n \geq 2$;

see Fortini et al. (2000, Theorem 3.1). Notably, the de Finetti measure μ in (1) can be recovered as the limiting law of the predictive distributions P_n. In fact, as shown by Berti and Rigo (1997), $\sup_x |P_n(x|x_{1:x}) - \hat{F}_n(x)| \to 0$ P-a.s., so that $\tilde{F}$ corresponds to the weak limit of P_n.

Under some additional conditions on P_n, the support of μ in (1) can be restricted to the set of absolutely continuous distributions. Let λ be the Lebesgue measure on $\mathbb{R}$ and λ^n be the n-product measure. According to Berti et al. (2013, Theorem 1), the random probability $\tilde{F}$ in (1) is absolutely continuous with respect to λ, write $\tilde{F} \ll \lambda$, if and only if:

(i) the finite-dimensional distributions of $X_1, \ldots, X_n$ are absolutely continuous with respect to λ^n for all n;
(ii) $\|P_n - \tilde{F}\|_{TV} \to 0$ P-a.s., where $\|\cdot\|_{TV}$ is the total variation norm.

Condition (i) amounts to $P_n \ll \lambda$, so we denote by m_n the predictive density according to $\mathrm{d}P_n(x|x_{1:n}) = m_n(x|x_{1:n})\, \mathrm{d}\lambda(x)$. It turns out that an exchangeable law P with absolutely continuous $\tilde{F}$ can be constructed starting from the sequence of predictive densities $(m_n)_{n \geq 1}$. In fact, according to Berti et al. (2013, Theorem 4), under a uniform integrability condition on m_n, $\|P_n - \tilde{F}\|_{TV} \to 0$ P-a.s. and, in turns, $\tilde{F} \ll \lambda$. If we denote by $\tilde{f}$ the random density associated to $\tilde{F}$ and by $m(x_1, \ldots, x_n)$ the density of $X_1, \ldots, X_n$ corresponding to $(m_n)_{n \geq 1}$, we have the following representation theorem:

$$m(x_1, \ldots, x_n) = \int_{\mathcal{F}_0} \prod_{i=1}^{n} f(x_i)\, \mathrm{d}\mu(f) \tag{2}$$

where μ now denotes the probability distribution of $\tilde{f}$ on $\mathcal{F}_0$, the space of density functions on $\mathbb{R}$. Moreover, since the total variation norm corresponds to the L_1-norm $\|\cdot\|_1$ in $\mathcal{F}_0$, (ii) implies

$$\|m_n - \tilde{f}\|_1 \to 0, \quad P\text{-a.s.} \tag{3}$$

In summary, a sequence of predictive densities $(m_n)_{n \geq 1}$, satisfying conditions (a) and (b) and the uniform integrability condition of Berti et al. (2013, Theorem 4), yields a continuous Bayesian model with de Finetti measure determined by the limit probability law of m_n.

A parametric model can be now characterized in terms of a *predictive sufficient* statistic, that is a random quantity $T = T(X_1, \ldots, X_n)$ with values in $\mathbb{R}^d$ which satisfies

$$P(X_{n+1} \in \cdot\, | X_1, \ldots, X_n) = P(X_{n+1} \in \cdot\, | T(X_1, \ldots, X_n)), \quad P\text{-a.s.}$$

By exchangeability, the empirical distribution $\hat{F}_n$ is a predictive sufficient statistic, hence we can write T as a function of $\hat{F}_n$, $T = T(\hat{F}_n)$. By de Finetti representation theorem, $P_n(\cdot|X_1, \ldots, X_n) = P_n(\cdot|T(\hat{F}_n))$ converges weakly to $\tilde{F}$, P-a.s. In the continuous case we write $m_n(x|t)$ for the conditional density of X_{n+1} given $T(\hat{F}_n) = t$. According to Fortini et al. (2000, Theorem 7.1), $T(\hat{F}_n)$ converges weakly, P-a.s., to a random element $\tilde{\theta}$ with value in $\Theta \subseteq \mathbb{R}^d$ under a regularity condition on $m_n(x|t)$ which corresponds to continuity of $m_n(x|t)$ in t uniformly in n. Then, representation (2) takes form

$$m(x_1, \ldots, x_n) = \int_{\Theta} \prod_{i=1}^{n} f(x_i; \theta) \pi(\theta)\, \mathrm{d}\theta \tag{4}$$

where $\pi(\theta)$ is the density of $\tilde{\theta}$ and $f(x; \theta)$ is the limit form of the predictive density $m_n(x|t)$. See Fortini et al. (2000) for examples of predictive characterization of classical parametric models. We can now identify the statistical model with $\{f(x; \theta), \theta \in \Theta\}$ and the prior distribution with $\pi(\theta)$, the latter seen as the limit law of the predictive sufficient statistic $T(\hat{F}_n)$. Bayes theorem is now applied to derive the posterior $\pi_n(\theta)$, that is the conditional density of $\tilde{\theta}$ given $X_1, \ldots, X_n$

$$\pi_n(\theta) = \frac{\prod_{i=1}^{n} f(X_i; \theta) \pi(\theta)}{\int_{\Theta} \prod_{i=1}^{n} f(X_i; \theta) \pi(\theta)\, \mathrm{d}\theta}.$$

It has to be said that the predictive characterization of parametric models is mainly of theoretical interest. In practice, the predictive densities $(m_n)_{n \geq 1}$ are the end-product rather than the origin of a statistical model and a prior. According to Walker's view, one can separate the mathematical construction leading to the representation theorem (4) from the stochastic process generating the sequence $(X_n)_{n \geq 1}$, and consider $(m_n)_{n \geq 1}$ as a learning scheme, or a "sequences of guesses", which justifies Bayes theorem only through (4), i.e. irrespectively of the way the data are generated. In this sense, rather than viewed as the limit form of predictive densities, $f(x; \theta)$ is chosen on a different ground, for example as a suitable approximation of the true data generating density f_0. This however poses the problem of how to interpret the prior since θ cannot be seen anymore as a large-sample function of the observables. The answer provided by Walker is that $\pi(\theta)$ should convey information about the parameter value θ_0 that minimizes the Kullback–Leibler divergence relative to f_0, i.e.

$$\theta_0 = \arg \min_{\theta \in \Theta} \left\{ - \int \log f_\theta(x)\, f_0(x)\, \mathrm{d}x \right\}. \tag{5}$$

and its use through Bayes theorem has to be validated by showing that $\pi_n(\theta)$ accumulates at θ_0 as $n \to \infty$

$$\pi_n\{\theta : \ |\theta - \theta_0| > \epsilon\} \to 0 \tag{6}$$

Please cite this article as: De Blasi, P., Discussion on article "Bayesian inference with misspecified models" by Stephen G. Walker. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.05.015

in P_0^n probability, where P_0^n is the n-product probability measure associated to iid sampling from f_0. A key point is that, unlike in limiting arguments like (3), convergence in (6) is not with respect to the exchangeable law P. Hence the need of referring to asymptotic evaluation like (6) as "frequentist" asymptotics, in opposition to the type of asymptotics considered in the predictive model representation. Indeed, since the seminal work by Diaconis and Freedman (1986), posterior asymptotics has been investigated exclusively in this frequentist setting.

2. Nonparametric estimation of the discrepancy

Given the acknowledgement of a misspecified model, the estimation of the discrepancy of a given parametric model with respect to the true data generating density is a worthy task. Let $(X_n)_{n\geq 1}$ be an iid sequence from a density f_0 and $\{f_\theta : \theta\in\Theta\}$ be a family of densities indexed by θ with prior $\pi(\theta)$. Note that we write $f_\theta(x)$ in place of $f(x;\theta)$ for notational convenience. Let θ_0 be defined in (5) as the parameter value that minimizes the Kullback–Leibler divergence relative to f_0, provided it exists and is unique. We measure the discrepancy of the parametric family with any divergence of the type

$$D(f_0,f_{\theta_0}) = \int f_{\theta_0} g[f_0(x)/f_{\theta_0}(x)]\,\mathrm{d}x$$

where g is a convex an positive function such that $g(1)=0$. See Liese and Vajda (2006). It is clear that we need to estimate the correction function $C_0(x)=f_0(x)/f_{\theta_0}(x)$ in order to compute $D(f_0,f_{\theta_0})$. In this section we consider a Bayesian nonparametric model built around $\{f_\theta : \theta\in\Theta\}$ where the interest is in estimating θ_0 and $C_0(x)$.

Let F_θ be the distribution function associated to f_θ and $p_Z(t)$ be a density on the unit interval depending on a random quantity Z (to be defined later) with prior $\Pi(\mathrm{d}Z)$. We define a density model through the probability transform

$$f_{\theta Z}(x) = p_Z(F_\theta(x))f_\theta(x), \quad x\in\mathbb{R} \tag{7}$$

which contains f_θ as special case when $p_Z(t)$ is the uniform density on $[0,1]$. See Verdinelli and Wasserman (1998) and Rousseau (2008) for applications in goodness-of fit testing. As for $p_Z(t)$, we set

$$p_Z(t) = \frac{\Psi(Z(t))}{\int_0^1 \Psi(Z(s))\,\mathrm{d}s}, \quad t\in[0,1]$$

where $Z(t)$ is a mean zero Gaussian process with covariance kernel $\sigma(s,t)$ and $\Psi(\cdot)$ is a cumulative distribution function with smooth unimodal symmetric density on $\mathbb{R}$. By the change of variable $s=F_\theta(x)$, the normalizing constant can be written as $\int_0^1 \Psi(Z(s))\,\mathrm{d}s = \int_{\mathbb{R}} \Psi(Z_\theta(x))f_\theta(x)\,\mathrm{d}x$ for any $\theta\in\Theta$, where $Z_\theta(x):=Z(F_\theta(x))$ is a mean zero Gaussian process with covariance $\sigma(F_\theta(\cdot),F_\theta(\cdot))$. Hence we can write

$$f_{\theta Z}(x) = \frac{\Psi(Z_\theta(x))f_\theta(x)}{\int_{\mathbb{R}}\Psi(Z_\theta(x))f_\theta(x)\,\mathrm{d}x}, \quad x\in\mathbb{R}.$$

and see $f_{\theta Z}(x)$ as $f_\theta(x)$ perturbed by $\Psi(Z_\theta(x))$. See Lenk (2003) and De Blasi and Walker (in press) for similar ideas. For given θ

$$C(x;\theta,Z) = p_Z(F_\theta(x)) = \Psi(Z_\theta(x)) \Big/ \int_{\mathbb{R}} \Psi(Z_\theta(x))f_\theta(x)\,\mathrm{d}x$$

describes the correction function f_0/f_θ and is considered the infinite-dimensional parameter of interest. Bayesian inference on $C(x;\theta,Z)$ for θ fixed can be based on the nonparametric model $\{f_{\theta,Z},\Pi(\mathrm{d}Z)\}$ via the posterior distribution

$$\Pi(\mathrm{d}Z|\theta,x_1,\ldots,x_n)\propto\Pi(\mathrm{d}Z)\times\prod_{i=1}^n f_{\theta Z}(x_i).$$

It turns out that the nonparametric model is flexible enough to recover the true density f_0 for any θ. Assume that f_0 is continuous and positive on all $\mathbb{R}$ and that f_θ satisfies $\lim_{x\to\pm\infty} f_0(x)/f_\theta(x)=0$. Let also $\log\Psi(u)$ be a Lipschitz function on $\mathbb{R}$. Denote by $\mathcal{A}(\sigma)$ the reproducing kernel Hilbert space of the Gaussian process Z and by $\overline{\mathcal{A}}(\sigma)$ its closure with respect to the sup norm on $[0,1]$. See van der Vaart and van Zanten (2008) for a formal definition. It can be shown that

$$\Pi\{Z:\ \|f_0-f_{\theta Z}\|_1>\epsilon|\theta,X_1,\ldots,X_n\}\to 0, \tag{8}$$

in P_0^n-probability as $n\to\infty$ provided that $\overline{\mathcal{A}}(\sigma)$ contains any continuous functions on $[0,1]$. The proof consists in the verification of an entropy condition on the space $\mathcal{F}_0$ and a prior support condition known as Kullback–Leibler property, see Ghosal et al. (1999, Theorem 2). As for the entropy condition, one can use van der Vaart and van Zanten (2008, Theorem 2.1) by establishing an appropriate relation between the Hellinger distance among densities of form (7) and the sup distance in the space of real-valued functions on $[0,1]$. See De Blasi and Walker (in press) for similar arguments. As for the Kullback–Leibler property, one can prove that the map $s\mapsto\Psi^{-1}[f_0(F_\theta^{-1}(s))/Mf_\theta(F_\theta^{-1}(s))]$ is well approximated by a function in $\overline{\mathcal{A}}(\sigma)$ for M a large enough positive constant, see Tokdar et al. (2010, Theroerm 3.1) for similar arguments. Note that posterior consistency in (8) implies that

$$C(x;\theta,x_1,\ldots,x_n) = \int C(x;\theta,Z)\Pi(\mathrm{d}Z|\theta,x_1,\ldots,x_n) \tag{9}$$

Please cite this article as: De Blasi, P., Discussion on article "Bayesian inference with misspecified models" by Stephen G. Walker. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.05.015

is a consistent estimate of the correction function $C_\theta := f_0/f_\theta$ in the L_1-integrated topology since (8) can be written as $\Pi\{Z: \int_{\mathbb{R}} |C_\theta(x) - C(x,\theta,Z)| f_\theta(x)\,dx > \epsilon | \theta, X_1, \ldots, X_n\} \to 0$.

Consider now the semi-parametric model $\{f_{\theta,Z}, \Pi(dZ), \pi(\theta)\}$. It is clear that $f_{\theta,Z}$ is an over-parametrized density: for any θ there is a Z such that $f_0 = f_{\theta,Z}$. Because of the lack of identification, the Bayes posterior $\Pi_n(d\theta, dZ) \propto \pi(\theta) d\theta \Pi(dZ) \times \prod_{i=1}^n f_{\theta,Z}(X_i)$ is not appropriate for estimating $C(x;\theta,Z)$ as we are interested to learn about a quantity C_0 which depends on a particular value of θ, i.e. θ_0. In De Blasi and Walker (in press) it is argued that one should use a different updating scheme for (θ, Z), namely

$$\tilde{\Pi}_n(d\theta, dZ) := \Pi(dZ|\theta, X_1, \ldots, X_n) \times \pi_n(\theta)\,d\theta \tag{10}$$

where $\pi_n(\theta) \propto \pi(\theta) \prod_{i=1}^n f_\theta(X_i)$ is the parametric posterior of θ. The joint distribution (10) can be justified in terms of estimating the posterior mean $C(x;\theta,x_1,\ldots,x_n)$ in (9) with respect to the parametric model $\{f_\theta(x), \pi(\theta)\}$, when the former is seen as a functional of θ and the data. It also corresponds to modifying the conditional posterior of θ of the semi-parametric model so to prevent estimation of θ to be confounded by estimation of Z. A standard result on misspecified parametric models is that π_n accumulates its mass at θ_0 with rate $1/\sqrt{n}$

$$\pi_n\{\theta: |\theta - \theta_0| > M_n n^{-1/2}|\} \to 0, \tag{11}$$

in P_0^n-probability for any sequence $M_n \to \infty$, see Kleijn and van der Vaart (2012, Theorem 3.1). Therefore it is worth exploring whether, in view of (8), $\tilde{\Pi}_n$ accumulates at C_0 in the L_1-integrated topology

$$\tilde{\Pi}_n\left\{(\theta,Z); \int |C_0(x) - C(x;\theta,Z)| f_{\theta_0}(x)\,dx > \epsilon\right\} \to 0, \tag{12}$$

in P_0^n-probability as $n \to \infty$. Results (11) and (12) would then provide an asymptotic validation of Bayesian updating (10).

References

Berti, P., Rigo, P., 1997. A Glivenko–Cantelli theorem for exchangeable random variables. Statistics & Probability Letters 32, 385–391.
Berti, P., Pratelli, L., Rigo, P., 2013. Exchangeable sequences driven by an absolutely continuous random measure. Annals of Probability, 41, 2090–2102.
De Blasi, P., Walker, S.G. Bayesian estimation of the discrepancy with misspecified parametric models. Bayesian Analysis, in press.
Diaconis, P., Freedman, D., 1986. On the consistency of Bayes estimates. Annals of Statistics 14, 1–26.
Fortini, S., Ladelli, L., Regazzini, E., 2000. Exchangeability, predictive distributions and parametric models. Sankhya, Series A 62, 86–109.
Ghosal, S., Ghosh, J.K., Ramamoorthi, A., 1999. Posterior consistency of Dirichlet mixtures in density estimation. Annals of Statistics 27, 143–158.
Kleijn, B.J.K., van der Vaart, A.W., 2012. The Bernstein–Von Mises theorem under misspecification. Electronic Journal of Statistics 6, 354–381.
Liese, F., Vajda, I., 2006. On divergence and informations in statistics and information theory. IEEE Transactions on Information Theory 52, 4394–4412.
Lenk, P.J., 2003. Bayesian semiparametric density estimation and model verification using a logistic Gaussian process. Journal of Computational and Graphical Statistics 12, 548–565.
Rousseau, J., 2008. Approximating interval hypothesis: p-values and Bayes factors. Bayesian Statistics 8, 417–452.
Tokdar, S.T., Zhu, Y.M., Ghosh, J.K., 2010. Bayesian density regression with logistic Gaussian process and subspace projection. Bayesian Analysis 5, 319–344.
van der Vaart, A.W., van Zanten, J.H., 2008. Rates of contraction of posterior distributions based on Gaussian process priors. Annals of Statistics 36, 1435–1463.
Verdinelli, I., Wasserman, L., 1998. Bayesian goodness-of-fit testing using infinite-dimensional exponential families. Annals of Statistics 26, 1215–1241.

Please cite this article as: De Blasi, P., Discussion on article “Bayesian inference with misspecified models” by Stephen G. Walker. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.05.015